Malaysian Journal of Applied Sciences 2016, **Vol 1(1)**: 45-54
©Universiti Sultan Zainal Abidin
eISSN 0127-9246 (Online)

Malaysian Journal of Applied Sciences

ORIGINAL ARTICLE

A New Strategy of Handling General Insurance Modelling Using Applied Linear Method

Wan Muhamad Amir W Ahmad [a], Mohamad Arif Awang Nawi [b], and Mustafa Mamat [b]

[a] School of Dental Sciences, Universiti Sains Malaysia, Health Campus, 16150 Kubang Kerian, Kelantan, Malaysia
[b] Faculty Informatics and Computing, Universiti Sultan Zainal Abidin, Tembila Campus, 22200 Besut, Terengganu, Malaysia

*Corresponding author: wmamir@usm.my

Received: 04/12/2015, Accepted: 17/03/2016

Abstract

This paper proposes the use of bootstrap, robust and fuzzy multiple linear regressions method in handling general insurance in order to get improved results. The main objective of bootstrapping is to estimate the distribution of an estimator or test statistic by resampling one's data or a model estimated from the data under conditions that hold in a wide variety of econometric applications. In addition, bootstrap also provides approximations to distributions of statistics, coverage probabilities of confidence intervals, and rejection probabilities of hypothesis tests that produce accurate results. In this paper, we emphasize the combining and modelling using bootstrapping, robust and fuzzy regression methodology. The results show that alternative methods produce better results than multiple linear regressions (MLR) model.

Keywords: Multiple linear regression; MM estimation; robust regression; bootstrap method; fuzzy regression

Introduction

Multiple linear regression modelling is a very powerful technique in statistics and is widely used in numerous research fields including finance, economic, agriculture. This method estimates linear relationship between dependent (response) and independent (explanatory) variables. The multiple linear regression model is expressed as $Y = b_0 + b_1 X_1 + \ldots + b_n X_n + \ell$ where b 's is parameters and ℓ is the error term assumed to be, following a normal distribution. The parameters are usually estimated using method of least squares. A good explanation of various aspects of multiple linear regression methodology is given in Draper and Smith (1998).

The primary goal of robust regression is to provide resistant results in the presence of outliers. In pursuit of this stability, robust regression limits the influence of outliers. Robust regression analysis provides an alternative to the least squares regression when fundamental assumptions are unfulfilled by the nature of the data (Marona et al., 2006). The properties of efficiency, breakdown, and high leverage points are used to define robust techniques

performance in a theoretical sense. One of the goals of robust estimator is a high finite sample breakdown point defined by Donoho and Huber (1983). Christmann (1994) and Rousseeuw and Leroy (1987) state that the breakdown point could be defined as the point or limiting percentage of contamination in the data at which any test statistics first becomes swamped. Hence, the breakdown point is simply the initial point at which any statistical test becomes swamped due to contaminated data. Some regression estimators have the smallest possible breakdown point of 1/n or 0/n. In other words, only one outlier would cause the regression equation to be rendered useless. Other estimators have the highest possible breakdown point of n/2 or 50%. If robust estimation technique has a 50% breakdown point, then 50% of the data could contain outliers and the coefficients would remain useable.

MM estimation is a special type of M-estimation developed by Yohai (1987). In his paper, Stromberg (1993) states that MM-estimation is a combination of high breakdown value and efficient estimations. Yohai's MM estimator was the first estimation of a high breakdown point and high efficiency under normal error. MM-estimators have three-stage procedures;

1. The first stage involves the calculation of S-estimate with influence function $\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{\tau_i}{s}\right)=K$; K is a constant, objective function ρ satisfies as:
   i. ρ is symmetric and continuously differentiable, and $\rho(0)=0$.
   ii. There exists $\alpha>0$ such that ρ is strictly increasing on $[0,\alpha]$ and constant on $[\alpha,\infty)$.
   iii. $\frac{K}{\rho(\alpha)}=\frac{1}{2}$.

2. The second stage involves the calculation of MM parameters that provide the minimum value of $\sum_{i=1}^{n}\rho\left(\frac{y_i-x_i'\hat{\beta}_{MM}}{\hat{\sigma}_0}\right)$ where $\rho(x)$ is the influence function used in the first stage and $\hat{\sigma}_0$ is the estimate of scale form the first step (standard deviation of the residuals).

3. The final step computes the MM estimate of scale as the solution to $\frac{1}{n-p}\sum_{i=1}^{n}\rho\left(\frac{y_i-x_i'\hat{\beta}}{s}\right)=0.5$

Bootstrap is a technique for resampling based on random sorts with retrieval in the data forming a sample. Additionally, this method provides approximations to distributions of statistics, coverage probabilities of confidence intervals, and rejection probabilities of hypothesis tests that produce accurate results (Hall, 1992; Efron and Tibshyrani, 1993). The theoretical bootstrap model is as follows;

$$Y^{*}=X\hat{\beta}+u^{*} \tag{1}$$

where u^{*} is a random term obtained from the residuals $\hat{u}$ of the initial regression. At each iteration $b(b=1,....,B)$, a sample $\{y_i^{*}\}_{i=1}^{n}$ of size (n, 1), is created from the theoretical bootstrap model.

Since the OLS residuals are smaller than the errors they estimate, the random term of the theoretical bootstrap model is constructed from the following transform residuals which have the same norm as the error term u_i:

$$\tilde{u}_i = \frac{\hat{u}_i}{\sqrt{(1-h_i)}} - \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{u}_i}{\sqrt{(1-h_i)}}$$

The theoretical bootstrap model is hence expressed as:

$$y_i^*(b) = X_i\hat{\beta} + \tilde{u}_i^*(b), \; i = 1 \ldots n \tag{2}$$

where $\tilde{u}_i^*(b)$ is resampled from $\tilde{u}_i$. Let us consider the random variable z_j, defined as $z_j = \frac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)}$, the standard confidence interval of β_j derives from the assumption according to which z_j is distributed according to a student's distribution with *n-p* degrees of freedom. Thus for a confidence level $(1-2\alpha)$, this confidence interval takes the following form:

$$[\hat{\beta}_j - s(\hat{\beta}_j)\cdot t_{(1-\alpha),n-p}, \hat{\beta}_j - s(\hat{\beta}_j)\cdot t_{(\alpha),n-p}] \tag{3}$$

where t is the percentile values (α) and $(1-\alpha)$ from the bootstrap-t percentiles with $n-p$ degrees of freedom. The bootstrap confidence intervals are constructed from two percentile and percentile-t approaches. The first method, based exclusively on bootstrap estimations, is the simplest one for obtaining confidence intervals. For a level $(1-2\alpha)$, the percentile confidence interval for parameter β_j is given by:

$$\left[\hat{\beta}_j^*(\alpha B), \hat{\beta}_j^*((1-\alpha)B)\right] \tag{4}$$

where $\hat{\beta}_j^*(\alpha B)$ is the αB -th value (respectively $\hat{\beta}_j^*((1-\alpha)B)$ the $(1-\alpha)B$ -th value) of the ordered list of the *B* bootstrap replications. The threshold values are hence selected so that α% of the replications provide smaller (larger) $\hat{\beta}_j^*$ than the lower (upper) bound of the percentile confidence interval.

A fuzzy regression model corresponding to multiple linear regression equation could be stated as;

$$y = A_0 + A_1x_1 + A_2x_2 + \ldots + A_kx_k \tag{5}$$

Previously, explanation variables $x_i's$ are assumed to be precise. However, according to the equation above, response variable ***Y*** is not crisp but is instead fuzzy in nature. That means the parameters are also fuzzy in nature. Our objective is to estimate these parameters. In further discussion, $A_i's$ are assumed as symmetric fuzzy numbers which could be presented by interval. For example, A_i could be expressed as fuzzy set given by $A_i =< a_{1c}, a_{1w} >$ where a_{ic} is centre and a_{iw} is radius or vagueness associated. Fuzzy set above reflects the confidence in the regression coefficients around a_{ic} in terms of symmetric triangular memberships function. Application of this method should be given more attention when the underlying phenomenon or the response variable is fuzzy. So, the relationship is also considered to be

fuzzy. This $A_i =< a_{1c}, a_{1w} >$ could be written as $A_1 = [a_{1L}, a_{1R}]$ with $a_{1L} = a_{1c} - a_{1w}$ and $a_{1R} = a_{1c} - a_{1w}$ (Kacprzyk and Fedrizzi, 1992). In fuzzy regression methodology, parameters are estimated by minimizing total vagueness in the model.

$$y_j = A_0 + A_1 x_{1j} + A_2 x_{2j} + \ldots + A_k x_{kj} \tag{6}$$

Using $A_i =< a_{1c}, a_{1w} >$, we could write

$$y_j =< a_{0c}, a_{0w} > + < a_{1c}, a_{1w} > x_{1j} + \ldots + < a_{nc}, a_{nw} > x_{nj}$$
$$=< a_{jc}, a_{jw} > .$$

Thus $y_{jc} = a_{0c} + a_{1c} x_{1j} + \cdots + a_{nc} x_{nj}$

$y_{jw} = a_{0w} + a_{1w} |x_{1j}| + \cdots + a_{nw} |x_{nj}|$

As y_{jw} represent radius and could not be negative, therefore on the right-hand side of equation $y_{jw} = a_{0w} + a_{1w} |x_{1j}| + \cdots + a_{nw} |x_{nj}|$, absolute values of x_{ij} are taken. Suppose there are *m* data point, each comprising $a(n+1) - row$ vector. Then parameters A_i are estimated by minimizing the quantity, which is total vagueness of the model-data set combination, subject to the constraint that each data point must fall within estimated value of response variable.

Materials and Methods

A Case Study of General Insurance

Table 1. Description of the variables

Variables	Description
Y	Profitability of General Insurance Companies
X_1	Net Investment Income
X_2	Total Liabilities and Assets
X_3	Management Expenses
X_4	Annual Premium
X_5	Net Claims Paid by The Company

Source: (Nawi, et al. 2012)

```
/* First we do Multiple linear regression */
procreg data= general;
model y=x1 x2 x3 x4 x5 ;
run;
```

Approach the MM-Estimation Procedure for Robust Regression

```
/* Then we do robust regression, in this case, MM-estimation */
ods graphics on;
procrobustreg method= MM fwls data= general plot=fitplot(nolimits)
plots=all;
model y = x1  x2 x3 x4 x5/ diagnostics itprint;
output out=resids out=robout r=residual weight=weight outlier=outlier
sr=stdres;
run;
ods graphics off;
```

Procedure for Bootstrap with Case Resampling (n =100)

```
/* And finally we use a bootstrap with case resampling */
ods listing close;
procsurveyselect data=general out=boot1 method=urs samprate=1outhits
rep=100;
run;
```

Procedure for Bootstrap into Fuzzy Regression Model

```
/*Combination of Bootstrap Technique with Fuzzy Regression*/
ods listing close;
procoptmodel;
set j= 1..30;
Number y{j}, x1{j}, x2{j}, x3{j}, x4{j}, x5{j};
read data boot1 into [_n_]  y x1 x2 x3 x4 x5;

/*Print y x1 x2 x3 x4 x5*/
Print y x1 x2 x3 x4 x5;
number n init 30;  /*Total of Observations*/

/* Decision Variables bounded or not bounded*/
/*Theses three variables are bounded*/
var aw{1..6}>=0;

/*These three variables are not bounded*/
var ac{1..6};

/* Objective Function*/
min z1= aw[1] * n + sum{i in j} x1[i] * aw[2]+sum{i in j} x2[i] *
aw[3]+sum{i in j} x3[i] * aw[4]+sum{i in j} x4[i] * aw[5]+sum{i in j} x5[i]
* aw[6];

/*Linear Constraints*/
con c{i in 1..n}:
ac[1]+x1[i]*ac[2]+x2[i]*ac[3]+x3[i]*ac[4]+x4[i]*ac[5]+x5[i]*ac[6]-aw[1]-
x1[i]*aw[2]-x2[i]*aw[3]-x3[i]*aw[4]-x4[i]*aw[5]-x5[i]*aw[6]<=y[i];

con c1{i in 1..n}:
ac[1]+x1[i]*ac[2]+x2[i]*ac[3]+x3[i]*ac[4]+x4[i]*ac[5]+x5[i]*ac[6]+aw[1]+x1[
i]*aw[2]+x2[i]*aw[3]+x3[i]*aw[4]+x4[i]*aw[5]+x5[i]*aw[6]>=y[i];

expand;/* This provides all equations */
solve;
print ac aw;
quit;
ods rtf close;
```

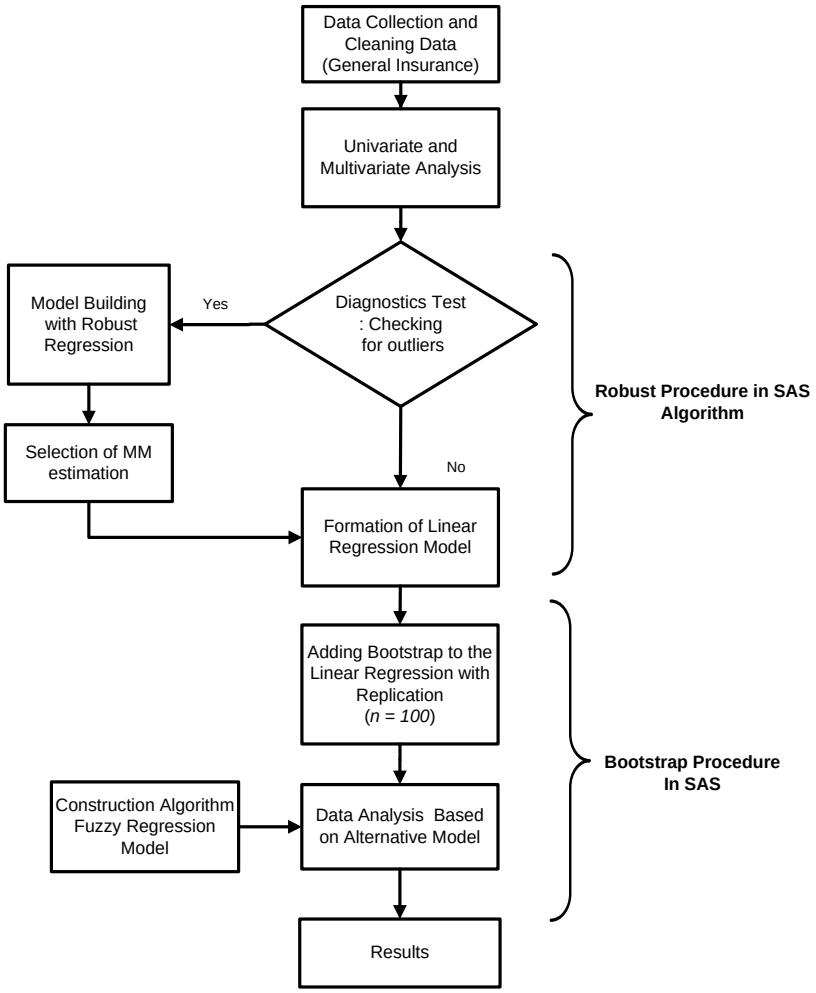

Figure 1. Flow chart of robust, bootstrap and fuzzy regression

Results and Discussion

A higher R-squared value shows how well the data fit the model and indicates a better model.

Table 2. Goodness-of-fit

Statistic	Value
R-Square	0.7764
AICR	2448.3810
BICR	2489.9700
Deviance	3.3609

Using the method of Multiple linear regression (MLR), we obtained the result as shown in Table 3 using bootstrapping method for fuzzy regression with n = 100. The aim of bootstrapping procedure is to approximate the entire sampling distribution of some estimator by resampling (simple random sampling with replacement) from the original data (Yaffee, 2002).

Table 3. Parameter estimates for final weighted least squares fit

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	0.2587	0.0166	0.2261	0.2913	241.7400	<.0001
X_1	1	0.0025	0.0038	-0.0049	0.0098	0.4200	0.5160
X_2	1	0.0030	0.0015	0.0001	0.0060	4.0100	0.0452
X_3	1	0.0009	0.0037	-0.0063	0.0081	0.0600	0.8054
X_4	1	0.9220	0.0020	0.9180	0.9259	210024	<.0001
X_5	1	0.0579	0.0025	0.0531	0.0627	555.1900	<.0001
Scale	0	0.0311					

Method of Fuzzy Regression (FR) (OPTMODEL)

Table 4 summarizes the findings of the calculated parameter. When using bootstrap procedure, we generate different output while using AC or AW, where AC denotes the centre and AW denotes the radius, i.e. half of the width of A.

The next step is to compare the performance of multiple linear regression and fuzzy regression.

Table 4. Value of center (AC) and radius (AW)

	AC	AW
1	1.3069	0.0000
2	0.4240	0.0000
3	-0.0478	0.0000
4	-0.4355	0.0000
5	0.9180	0.0000
6	0.1637	0.0058

The Fitted Model for Multiple Linear Regressions

$$Y = 0.2587 + 0.0025\ X_1 + 0.0030\ X_2 + 0.0009\ X_3 + 0.9220\ X_4 + 0.0579\ X_5 \quad (7)$$

Standard Error (0.0166) (0.0038) (0.0015) (0.0037) (0.002) (0.0025)

The upper limits of prediction interval are computed by coefficient plus standard error

$$Y = (0.2587 + 0.0166) + (0.0025 + 0.0038)\ X_1 + (0.0030 + 0.0015)\ X_2 + (0.0009 + 0.0037) X_3 + (0.9220 + 0.0020) X_4 + (0.0579 + 0.0025) X_5$$

The lower limits of prediction interval are computed by coefficient minus standard error

$$Y = (0.2587 - 0.0166) + (0.0025 - 0.0038)\ X_1 + (0.0030 - 0.0015)\ X_2 + (0.0009 - 0.0037) X_3 + (0.9220 - 0.0020) X_4 + (0.0579 - 0.0025) X_5$$

Table 5. Average width for former multiple linear regression model and fuzzy bootstrap regression model

Multiple Linear Regression Model			**Fuzzy Bootstrap Regression Model**		
Lower Limit	**Upper Limit**	**Width**	**Lower Limit**	**Upper Limit**	**Width**
11.41	11.73	0.32	11.46	11.58	0.12
12.02	12.37	0.35	11.80	11.93	0.13
12.18	12.53	0.35	12.31	12.44	0.14
12.11	12.47	0.36	12.27	12.41	0.14
12.44	12.79	0.35	12.58	12.72	0.14
10.99	11.31	0.32	11.05	11.17	0.12
11.85	12.19	0.34	11.97	12.10	0.13
13.67	14.06	0.39	13.95	14.11	0.16
12.40	12.75	0.35	12.69	12.82	0.13
12.34	12.69	0.36	12.46	12.61	0.14
11.34	11.67	0.33	11.41	11.54	0.12
11.10	11.43	0.32	11.09	11.22	0.12
11.80	12.14	0.34	11.95	12.08	0.13
11.43	11.78	0.35	11.49	11.63	0.13
12.11	12.46	0.35	12.26	12.40	0.14
11.42	11.75	0.33	11.55	11.68	0.13
10.64	10.96	0.32	10.76	10.88	0.13
9.55	9.84	0.29	9.65	9.75	0.10
11.07	11.40	0.33	11.19	11.31	0.12
11.73	12.06	0.34	11.85	11.98	0.13
11.78	12.13	0.34	11.85	11.98	0.13
11.36	11.70	0.34	11.51	11.64	0.14
12.58	12.94	0.36	12.63	12.78	0.14
12.47	12.82	0.36	12.63	12.77	0.14
11.54	11.86	0.33	11.61	11.73	0.12
13.12	13.49	0.37	13.17	13.31	0.14
12.31	12.66	0.35	12.42	12.56	0.14
12.58	12.94	0.36	12.65	12.79	0.14
12.18	12.54	0.36	12.29	12.43	0.14
12.12	12.47	0.35	12.24	12.38	0.14
Average		0.34	**Average**		0.13

The Fitted Model for Fuzzy Bootstrap Regression

$$Y = 1.3070 + 0.4240\ X_1 - 0.0478\ X_2 - 0.4355\ X_3 + 0.9180\ X_4 + 0.1637\ X_5 \quad (8)$$

The upper limits of prediction interval are computed by coefficient plus standard error

$$Y = [1.3070 + 0] + [0.4240 + 0]\ X_1 + [-0.0478 + 0]\ X_2 + [-0.4355 + 0] X_3 + [0.9180 + 0] X_4 + [0.1637 + 0.0058] X_5$$

The lower limits of prediction interval are computed by coefficient minus standard error

$$Y = [1.3070 - 0] + [0.4240 - 0]\ X_1 + [-0.0478 - 0]\ X_2 + [-0.4355 - 0] X_3 + [0.9180 - 0] X_4 + [0.1637 - 0.0058] X_5$$

The width of prediction intervals in respect of multiple linear regression model and fuzzy regression model corresponding to each set of observed explanatory variables were computed manually.

As shown in Table 5, the average width for former multiple regression was found to be 0.34 while using fuzzy regression, while the average width for fuzzy regression is 0.13 which indicates the superiority of fuzzy regression methodology. From this analysis, the most efficient method to obtained relationship between response and explanatory variable is to apply fuzzy regression method compared to linear regression method.

Conclusion

This paper discusses the combination of an algorithm with robust, fuzzy regression and bootstrap method. The reasons for using a small sample size were (a) to apply a bootstrap method in order to achieve an adequate sample size; (b) to compare the efficiency of original method and the bootstrap method; and (c) to give a better understanding on how the algorithm works. According to general insurance data, three independent variables in this case were significant to the profitability of general insurance companies. Without using robust and bootstrapping, the result shows that only one out of five variables were significant. Interestingly when using robust to detect outliers and to provide resistant results in the presence of outliers and bootstrapping method (with n = 100), the entire significant variable are included in the model. This algorithm provides us with improved understanding of the modified method and underlying relative contributions. Further study looking at possibility to approach response surface methodology for each of significant variables in single algorithm is warranted.

References

Christmann, A. (1994). Least median of weighted squares in logistic regression with large strata. *Biometrika, 81*, 413-417.

Donoho, D. L., & Huber, P. J. (1983). The notion of breakdown point. In Bickel P. J., Doksum K. A., & Hodges, J. L. (Eds.), *A festschrift for Erich, L. Lehmann* (pp. 157-184). Belmont: Wadsworth.

Draper, N., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York: Wiley.

Efron B., & Tibshyrani, R.J. (1993). *An introduction to the bootstrap*. New-York: Chapman and Hall.

Hall, P. (1992). *The bootstrap and edgeworth expansion*. New-York: Springer Verlag.

Kacprzyk, J., & Fedrizzi, M. (1992). *Fuzzy regression analysis*. Warsaw: Omnitech Press.

Marona, R., Martin, R., & Yohai, V. J. (2006). *Robust statistics theory and methods*. England: John Wiley & Sons Ltd.

Nawi, M. A. A., Ahmad, W. M. A. W., & Aleng, N. A. (2012). Efficiency of general insurance in Malaysia using stochastic frontier analysis (SFA). *International Journal of Modern Engineering Research, 2*(5), 3886-3890.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York: Wiley-Interscience.

Stromberg, A. J. (1993). Computation of high breakdown nonliner regression parameters. *Journal of the American Statistical Association, 88*(421), 237-244.

Yaffee, R. A. (2002). Robust Regression Analysis: Some Popular Statistical Package Options. *ITS Statistics, Social Science and Mapping Group, 23*, 1-12.

Yohai, V.J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics, 15*, 642-656.

How to cite this paper:
Ahmad, W.M.A.W., Nawi, M.A.A. & Mamat, M. (2016). A new strategy of handling general insurance modelling using applied linear method. *Malaysian Journal of Applied Sciences, 1*(1), 45-54.